REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		Final

4. TITLE AND SUBTITLE

Markov Dependence in Statistics and Information Theory, and Statistical Problems in Physical Mapping

5. FUNDING NUMBERS

DAAH04-94-G-0232

6. AUTHOR(S)

Dr. Bin Yu

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)

University of California-Berkeley
Berkeley, CA 94720

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

ARO 32320.8-MA

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

Investigations were carried out on Markov dependence in statistics and information theory, and statistical problems in physical mapping. Results were obtained on the Minimum Description Length Principle and statistical inference, adaptive quantization in image compression, Markov chain Monte Carlo methods, and statistical problems in the Human Genome Project. These results shed light on the connections between information theory and statistics, on the role of parametric models in quantization and image compression, on understanding the convergence behaviors of Markov chain Monte Carlo samplers, and on the information needed for a clone map of chromosomes. Furthermore, a wavelet image coder is designed as part of the investigation and it gives an excellent performance on test images.

14. SUBJECT TERMS			15. NUMBER IF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

ARO proposal number: P-32320-MA

Final Report:: 6/1/94 - 5/31/98

Title: Markov Dependence in Statistics and Information Theory and Statistical Problems in Physical Mapping

Grant number: DAAH04-94-G-0232

PI: Professor Bin Yu, Department of Statistics, University of California, Berkeley

(4) Statement of Problems Studied:

Investigations were carried out on Markov dependence in statistics and information theory, and statistical problems in physical mapping. Results were obtained on the Minimum Description Length Principle and statistical inference, adaptive quantization in image compression, Markov chain Monte Carlo methods, and statistical problems in the Human Genome Project. These results shed light on the connections between information theory and statistics, on the role of parametric models in quantization and image compression, on understanding the convergence behaviors of Markov chain Monte Carlo samplers, and on the information needed for a clone map of chromosomes. Furthermore, a wavelet image coder is designed as part of the investigation and it gives an excellent performance on test images.

(5) Summary of The most Important Results

For many years, the PI has been intrigued by the interplay between information/communication theory and statistics. In the period supported by the grant, the PI worked on Minimum Description Length (MDL) Principle of Rissanen, focusing on understanding MDL in comparison to more conventional statistical methods. Minimax lower bounds for smooth continuous Markov source classes are technically much more challenging to obtain than in the iid case. Nevertheless, Yu (1994) gives a minimax redundancy lower bound in the Markov case via a recursive equation. Based on mutual information calculations, Yu (1996) derives minimax redundancy lower bounds for smooth density classes and hence unifies the minimax approaches to redundancy lower bounds in the parametric and nonparametric cases. Yu (1997) explores connections between important inequalities in statistics and information theory. Rissanen and Yu (1995) study MDL in the computational learning context. Barron, Yang, and Yu (1994) show for the first time that the extra log(n) factor is not necessary for MDL-based density estimators. In particular, they constructed an optimal-rate MDL-based histogram estimator that takes into the account the Lipschitz condition in the coding of histogram parameters. Barron, Rissanen and Yu (1998) is an invited paper for the 50th anniversary of Information Theory to appear in a special issue of IEEE Transactions on Information Theory. It reviews the theoretical developments of MDL and makes connections among different formulations of the redundancy problem in coding which motivates and validates MDL. Moreover, explicit connections between statistical estimation and lossless data compression are also made.

Markov Chain Monte Carlo (MCMC) methods have attracted much attention from researchers as an important computational tool for a variety of applications including likelihood computation in frequentist statistics and posterior computation in Bayesian statistics. Because the target distribution is the stationary distribution for the constructed Markov chain, the success of the MCMC method relies crucially on our ability to assess the convergence of the chain to its equilibrium. The PI's interest in Markov Chain Monte Carlo (MCMC) has been on output error assessment and convergence diagnostics issues. Mykland, Tierney and Yu (1995) propose to use the split-chain idea of Nummelin, Athreya and Ney. In this way, known results for regenerative

simulations apply and more reliable estimates of the variance of the sample mean can be obtained from the simulated chain. In the discussion (Yu, 1995) on the Besag et al. paper in Statistical Science (cf. Yu and Mykland, 1998), the cusum plot is proposed as a simple diagnostic tool based on a one-dimensional summary of the MCMC sample. Strong approximation results for absolutely regular sequences are used to argue that the smoothness of the line-joined cusum path reflects the mixing speed of the one-dimensional summary statistic: the faster the mixing, the more ``hairy" the cusum plot. Ostland and Yu (1997) present a manually-adaptive extension of Quasi Monte Carlo (QMC) methods for approximating marginal densities as a viable alternative to the Metropolis algorithm - when the joint density is known up to a normalization constant. Randomization and a batch-wise approach involving (0,s)-sequences are the cornerstones of our method. By incorporating a variety of graphical diagnostics the method allows the user to adaptively allocate points for joint density function evaluations and therefore produces reliable marginal density approximations in moderate dimensions.

Much of the recent work on wavelet subband image coding has relied on adaptive quantization to achieve substantial gains over other traditional techniques - for example, those based on the Discrete Cosine Transform. In a typical example, different quantizers are used for different subbands, or for blocks within a given subband, and the quantizers are explicitly sent to the decoder as overhead. Then, the quantized coefficients are transmitted using adaptive entropy coding, typically through arithmetic coding. In this example, forward adaptation is used for the quantizers and backward adaptation for the entropy coding. Yoo, Ortega, and Yu (1997) show that a combination of forward and backward adaptation methods can be used to update the quantizers as well. Specifically, we propose an algorithm where classification can be done based on the quantized past data and where the quantizer to be used within one class can itself be adapted on the fly. Our proposed algorithm gives very competitive performance on standard test images (in terms of Signal to Noise ratio under the mean squared distortion measure).

The first step in the Human Genome Project is to assemble DNA fragments, called clones, to form clone maps, which allow the detailed study of chromosomal regions of biological interest. Yu and Speed (1997) answer biologist Lehrach's question about how much information is needed to complete a clone map by formulating a number of different notions (or configuration variables) of a clone map. The entropy of each notion (or configuration variable) is tightly bounded. The results are useful for planning future mapping efforts. In particular, based on the bounds in Yu and Speed (1997), comparisons are made for four ``model" species in terms of information needed for the mapping of their respective cosmid clone libraries. It follows that the cosmid clone mapping for the roundworm requires about 40 times as much information as that for the bacterium E. Coli, and that such mapping for humans requires about 1,500 times as much information as that for the bacterium E. coli. Another interesting fact which follows from the entropy bounds is that a variable relating to the pairwise approach to clone mapping contains a substantial proportion (more than 20%) of the information in a full configuration variable.

Various random fingerprinting methods are sometimes used to detect the overlap between pairs of clones as a first step towards producing a minimal tiling path of clones for subsequent mapping and sequencing efforts. Nelson, Speed and Yu (1997) analyze the overlap detection problem for two clones. They evaluate and compare various statistical procedures for the two-clone overlap based on random fingerprinting data. In particular, they quantify the limitations of random fingerprinting as a way to detect pairwise overlap, and within those limitations, the most effective ways to use the data. Based on the results, it is concluded that random fingerprinting based methods generate very weak overlap detectors, confirming what biologists are discovering by trial and error.

(6) **List of Publications and Technical Reports**

Barron, A. R. and Yang, Y. and Yu, B. (1994). Asymptotically optimal function estimation by minimum complexity criteria Proceedings of 1994 International Symposium on Information Theory, Trondheim, Norway, pp. 38.

Yu, B. (1994). Lower bound on the expected redundancy for classes of continuous Markov sources, In "Statistical Decision Theory and Related Topics V", S. S. Gupta and J. O. Berger (eds), pp. 453-466.

Yu, B. (1995). Comment: Extracting more diagnostic information from a single run using cusum path plot, Statistical Science, vol. 10, 54-58.

Rissanen, J. and Yu, B. (1995). MDL learning. In "Learning and Geometry: Computational Approaches, Progress in Computer Science and Applied Logic", David Kueker and Carl Smith (eds), pp. 3-19, Birkhauser, Boston.

Mykland, P. and Tierney, L. and Yu, B. (1995). Regeneration in Markov Chain sampler, Journal of American Statistical Association, vol. 90, pp. 233-241.

Yu, B. (1996). Lower bounds on expected redundancy for nonparametric classes, IEEE Trans. Information Theory, vol. 42, 272-275.

Yu, B. (1997). Assouad, Fano, and Le Cam. In "Festschrift for Lucien Le Cam", D. Pollard, E. Torgersen, and G. Yang (eds), pp. 423-435, Springer-Verlag.

Yu, B. and Speed, T. (1997). Information and the clone mapping of chromosomes. Annals of Statistics, vol. 25, 169-185.

Nelson, D. and Speed, T. and Yu, B. (1997) The limits of random fingerprinting, Genomics, vol. 40, 1-12.

Ostland, M. and Yu, B. (1997)
Exploring quasi Monte Carlo for marginal density approximation, Statistics and Computing, vol. 7, 217-228.

Yoo, Y. and Ortega, A. and Yu, B. (1997) Image subband coding using context-based classification and adaptive quantization. IEEE Trans. Image Proc. (revised).

Yu, B. and Mykland, P. (1998), Looking at Markov samplers through a simple diagnostic idea. Statistics and Computing, (in press).

Barron, A., Rissanen, J. and Yu, B. (1998). Minimum description length principle in coding and modeling. Invited paper for the 50th anniversary issue of IEEE. Trans. Inform. Th., Oct.

(7) **List of All Participating Scientific Personnel Showing Any Advanced Degrees Earned by Them While Employed on the project:**

Rebecka Jornsten, M.S., 1997